Robin Gras[1]
Markus Müller[1]
Elisabeth Gasteiger[1]
Steven Gay[1]
Pierre-Alain Binz[1,2]
William Bienvenut[2]
Christine Hoogland[1]
Jean-Charles Sanchez[2]
Amos Bairoch[1]
Denis F. Hochstrasser[2]
Ron D. Appel[1]

[1]Swiss Institute of Bioinformatics, University Medical Center, Geneva, Switzerland
[2]Central Clinical Chemistry Laboratory, Geneva University Hospital, Geneva, Switzerland

Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection

We have developed a new algorithm to identify proteins by means of peptide mass fingerprinting. Starting from the matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) spectra and environmental data such as species, isoelectric point and molecular weight, as well as chemical modifications or number of missed cleavages of a protein, the program performs a fully automated identification of the protein. The first step is a peak detection algorithm, which allows precise and fast determination of peptide masses, even if the peaks are of low intensity or they overlap. In the second step the masses and environmental data are used by the identification algorithm to search in protein sequence databases (SWISS-PROT and/or TrEMBL) for protein entries that match the input data. Consequently, a list of candidate proteins is selected from the database, and a score calculation provides a ranking according to the quality of the match. To define the most discriminating scoring calculation we analyzed the respective role of each parameter in two directions. The first one is based on filtering and exploratory effects, while the second direction focuses on the levels where the parameters intervene in the identification process. Thus, according to our analysis, all input parameters contribute to the score, however with different weights. Since it is difficult to estimate the weights in advance, they have been computed with a generic algorithm, using a training set of 91 protein spectra with their environmental data. We tested the resulting scoring calculation on a test set of ten proteins and compared the identification results with those of other peptide mass fingerprinting programs.

Keywords: Mass spectrometry / Peak detection / Peptide mass fingerprinting / Protein identification

EL 3747

1 Introduction

One of the tasks of proteomics is to identify the proteins expressed by an organism or tissue [1]. This requires several steps. The proteins are first isolated and some protein-specific attributes are measured. A protein sequence database is then screened in order to retrieve the protein or proteins that best match these attributes. Until recently, the attributes were most commonly determined by chemically extracting amino acid sequence information [2]. While these methods are reliable and can be fully-automated, they are slow and do not allow high throughput identification. Hence new techniques for protein identification had to be developed. A major impetus came from mass spectrometry of large molecules. New methods such as MALDI [3, 4] and electrospray ionization (ESI) [5], as well as new spectrometers [6] became available and made it possible to analyze proteins in small concentrations in a short time. Among the various spectrometric methods are: Fourier transform mass spectrometry (FTMS) [7], which provides a high mass resolution, and quadrupole time of flight (QTOF) [8], where ions of a small mass range are selected by a quadrupole ion trap and then transferred to a collision chamber before their fragments are analyzed. Furthermore, there are reflectron time-of-flight spectrometers (MALDI-TOF and ESI-TOF), which allow the measurement of masses in a large range with sufficient precision.

Currently the most common method to identify proteins is, first, to enzymatically digest the proteins, then to determine the masses of the resulting peptides by peak detection on a MALDI-TOF or ESI-TOF spectrum, and finalty to use the peptide mass fingerprints to search proteins sequence databases for correct matches. Optimizing the peak detection and database search algorithms is thus the key to improving protein identification from peptide mass fingerprints.

1.1 Peak detection

Peak detection is an important step in the identification process. Occasionally only a few experimental peptide masses in the fingerprint match the theoretical masses in

Correspondence: Dr. Robin Gras, Swiss Institute of Bioinformatics, 1. rue Michel-Servet, CH-1211 Geneva 4, Switzerland
E-mail: robin.gras@isb-sib.ch
Fax: +41-22-372-6198

0173-0835/99/1818-3535 $17.50+.50/0

a database; thus failure to detect one peak can hinder the correct identification of a protein. On the other hand, if too many false peaks are considered, this may lead to erroneous database matches. Furthermore, it is important to precisely determine the peptide masses. The peak detection algorithm must also be able to correct calibration errors of the mass spectrometers. Finally, the process of peak detection should be fast and fully-automated in order to grant high throughput data handling.

1.2 Identification

The principle of protein identification using peptide mass fingerprinting is based on the comparison of the list of experimental masses with a database containing the theoretical peptide masses of known proteins. The goal is to find the protein or proteins whose peptide masses provide the best match with the experimental fingerprint. It is worth mentioning that several other attributes of proteins may be useful in characterizing the likeness between the protein under investigation and identifying candidates from the database [9]. Information about the species, the molecular mass or the isoelectric point of the whole protein can be very helpful in selecting the right protein. Chemical modifications caused by biochemical mechanisms in the living cell or during the preparation of the experiment modify the peptide masses and also have to be taken into account while parsing the database.

Several programs exist that perform this kind of protein identification. They all use some of the available attributes and search various protein sequence databases. The critical question in this approach is to present the user with a ranking of the proteins that match the protein under investigation, which considerably facilitates the interpretation of the identification results. Most programs show scores associated with each protein, thus giving a degree of confidence in the matching protein. The simplest scoring method is to count the number of matching peptide masses. This is applied by the PeptideSearch program (http://www.mann.embl-heidelberg.de/Services/PeptideSearch/FR_PeptideSearchForm.html) which searches the nrdb database, as well as by the PepIdent program [10] (http://www.expasy.ch/tools/peptident.html) which searches the SWISS-PROT and TrEMBL databases [11]. In addition, PepIdent uses some of the annotations from SWISS-PROT to refine its search, taking into account known protein modifications (post-translational and processing of precursor molecules into mature chains and peptides). The Mowse program (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse) [12] determines a score by considering the frequency of each peptide mass in its database (OWL) in order to emphasize the rarest peptides. This score also takes into account the presence of missed cleavages in matched peptides, modifying their weight in the score by a fixed factor (pFactor). The MS-Fit program (http://prospector.ucsf.edu/ucsfhtml3.2/msfit.htm) uses the same scoring method on the (nonredundant) NCBInr database. ProFound [13] (http://prowl.rockefeller.edu/cgi-bin/ProFound) calculates a probability for the identification of the right protein, given by a bayesian formula, and uses the distance between experimental and theoretical masses obtained from the NCBInr database. Finally, the MassProfile program, included in the Darwin library (http://cbrg.inf.ethz.ch/) [14] also determined an identification score based on the probability of randomly obtaining a match of n experimental masses with n theoretical masses, given the interval of possible masses and the maximum allowed distance of masses accepted in this match.

All these algorithms utilize the various attributes presented above to control the number of proteins considered for the identification. However, they make little use of this information in their scoring calculation, since they use at most one or two of the attributes (distance between masses, presence of missed cleavages, mass distribution in the database, *etc*). These represent only a small part of the parameters that could influence the quality of identification. In order to better understand their respective role, we carried out a systematic study of the importance of each atttribute in the identification process. This led to the definition of a new scoring scheme that takes into account maximal information from each attribute, thus allowing for a better discrimination of candidate proteins and facilitating the identification of the right protein. This paper first presents an optimized automated peak detection algorithm and then details a new protein identification method, as well as its associated scoring procedure.

2 Materials and methods

2.1 Materials

2.1.1 Chemicals

SDS-PAGE molecular weight standards were purchased from Bio-Rad Laboratories (Hercules, CA, USA). Sequencing-grade modified trypsin was purchased from Promega (Madison, WI, USA). Trifluoroacetic acid (TFA) and α-cyano-4-hydroxy-*trans*-cinnamic acid (ACCA) were purchased from Sigma (St. Louis, MO, USA). Acetonitrile (AcCN), HPLC-grade, was purchased from Fluka (Buchs, Switzerland). Methanol (analytical grade) and sodium bicarbonate were purchased from Merck (Darmstadt, Germany). Immobilized pH gradient strips were purchased from Amersham-Pharmacia-Biotech (Uppsala, Sweden).

2.1.2 Sample preparation

Bio-Rad molecular weight standards were separated by 1-D SDS electrophoresis [15]. The other protein samples were separated by two-dimensional gel electrophoresis (2-DE) [16]. Gels were stained with Coomassie Brilliant Blue (CBB) R-250 (0.1% w/v), methanol (30% v/v) and acetic acid (10% v/v) for 30 min and were destained with repeated washes of methanol (40% v/v) and acetic acid (10% v/v) solution. Protein spots were excised and destained with 100 µL of a 50 mM ammonium bicarbonate solution at 37°C for 45 min. Destaining solution was removed and the gel pieces were dried under vacuum. The gel pieces were generally reswollen with 20 µL of 20 mM ammonium bicarbonate and 4 µL of 0.1 mg/mL of trypsin. The gel was dried to evaporate solvent and volatile salts, usually after overnight incubation at room temperature. Then 20 µL of 50% AcCN, 0.1% TFA were added for 10 min with sonication to extract peptides from the gel.

2.1.3 Mass spectrometry

Mass spectrometric measurements were performed on a MALDI-TOF mass spectrometer Voyager™ Elite (PerSeptive Biosystems, Framingham, MA, USA) equipped with 337 nm nitrogen laser. The analyzer was used in the reflectron mode at an accelerating voltage of 18–20 kV and a delayed extraction set to 100–140 ns. Laser power was generally set about 20% above threshold for matrix molecular ion production. Spectra were accumulated between 10–256 times. The matrix solution used was 4 mg/mL ACCA in 30–50% AcCN, 0.1% TFA.

2.1.4 Computer hardware

The programs are written in ANSI C++ and run on Unix and Windows systems. We also developed a Perl script that allows running the peak detection on a Windows PC and the database search on a Unix server.

2.2 Peak detection in MALDI-TOF mass spectra

2.2.1 Introduction

A MALDI-TOF spectrum is a sampled signal, *i.e.*, an array of floating point values that consists of trend, noise and peaks (Fig. 1). The trend or baseline is the signal produced by the electronics of the mass spectrometer that one would obtain if no material entered the mass spectrometer and in the absence of noise. It does not vary over small mass ranges (~ 10 Da). The noise, which is caused by electronic disturbances and fragments of material, varies over small mass ranges (< 1 Da) with little correlation, *i.e.*, each array value varies randomly and almost independently of its neighboring values. Peaks have a more or less predefined shape (see below) and are therefore strongly correlated. The notion of a peak may be misleading, because one "peak" actually consists

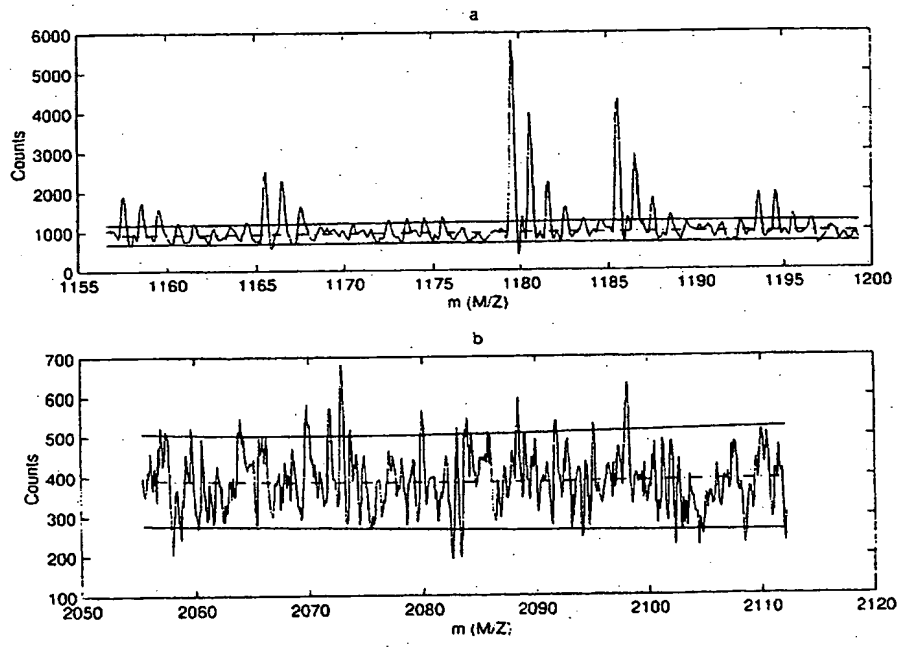

Figure 1. Mass spectrum of the *Escherichia coli* protein Prolyl-tRNA synthetase (SYP_ECOLI, P16659) digested with trypsin. The spectrum was acquired with a Voyager Elite MALDI-TOF mass spectrometer. Note that the mass unit is *M/Z*, where *M* is the mass in Da and *Z* the amount of unit charges of a peptide. (a) Part of the spectrum containing peaks; (b) a noisy region. The — line shows the trend, while the upper and lower solid lines show the trend plus and minus the noise, respectively.

of several so-called isotopic peaks [17] (Figs. 2 and 5). Determining the monoisotopic mass of peaks is a long existing task [18, 19], and software accomplishing this task is usually delivered together with the spectrometer hardware. Unfortunately, the software we had at hand lacked the necessary flexibility and accuracy. Therefore a custom peak detection program had to be conceived. It has been designed to yield a precise and fast localization of all the peaks, even if these are small or if several peaks overlap.

In order to detect peaks in a spectrum, we can apply a regression algorithm [20]. Let t_α be a template or model for a peak, where α is a set of parameters (such as height and width). A peak is detected in a spectrum if its template fits a part of that spectrum, *i.e.*, if a measure of distance between the spectrum and the template $|s-t_\alpha|$ has a local minimum and is smaller than a threshold value. The choice of t_α is crucial. Three conditions must be fulfilled: (i) the match should be clear (high signal-to-noise ratio), (ii) it should be precise (low deviation due to noise), and (iii) it should be unique (local minima not too close to each other). The first two conditions can be solved analytically. They yield $t_\alpha = p_\alpha$, where p_α denotes an ideal peak as it would appear in a noiseless spectrum. The third condition results in blurring the template t_α, *i.e.*, reducing its high frequency part and thereby smoothing the error function. Hence, it competes with the other two conditions. Canny [21] developed this theory for the case where only one pattern with a fixed shape is present. But as we will see below, we have to deal with peaks of variable shape, and the template has to adapt to these shapes.

2.2.2 Application to MALDI-TOF mass spectra

Since multiply charged peptides are rarely observed in MALDI-TOF spectra, we can assume that all peptides carry a single charge, thus the spacing between isotopic peaks is 1 Da. An isotopic distribution defines the probabilities that a molecule carries additional neutrons. For peptides of the same mass, there are several possible isotopic distributions. They depend on the atomic composition and particularly on the number of sulfur atoms [22]. However, the differences are scarcely visible in a mass spectrum, because the atomic compositions of peptides with the same mass are similar. Thus, we only consider an average isotopic distribution calculated by averaging all peptides with a mass in $[m, m+1]$ that are obtained from *in silico* digestion of the proteins in the SWISS-PROT database [22]. Let $p_m^{iso}(i)$, $i \in [0..\infty]$ denote the probability of having i additional neutrons in this average distribution. The probability of the monoisotopic peak $p_m^{iso}(0)$ decreases with increasing mass m (Fig. 2). This

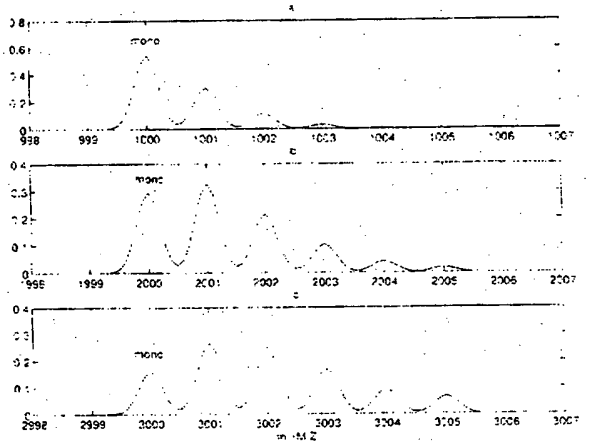

Figure 2. Average isotopic distributions $t_{m,h_o,h,\sigma}(m')$. The monoisotopic part is marked with the label 'mono'. (a) For $m = 1000$ Da; (b) $m = 2000$ Da; and (c) $m = 3000$ Da. $\sigma = 0.28$, $h_o = 0$ and $h = 1$.

feature is crucial for a correct peak detection in different peptide mass ranges.

A template can now be obtained from p_m^{iso}:

$$t_{\alpha}(m') = t_{m,h_o,h,\sigma}(m') = h_o - h\sum_i p_m^{iso}(i)\, e^{-\frac{(m'-m-i)^2}{2\sigma^2}} \qquad (1)$$

when h_o is the offset, h the height, m the monoisotopic mass and σ the width. The offset h_o is necessary to correct errors in the trend estimation (see Section 2.2.3).

Let us now define the error function e:

$$e(m) = \frac{1}{h^2(m_2 - m_1)} \min_{\alpha} \sum_{m_1}^{m_2} (s(m_i) - t_{\alpha}(m_i))^2 \qquad (2)$$

where m_i are the sampling values of the spectrum and $m_1 = 1$ and $m_2 = 5$ define the window in which Eq. (2) is evaluated. It must be large enough to contain the bulk of $t_{m,h_o,h,\sigma}(m')$. However, if it is too large, the $\Sigma s^2(m_i)$ term would dominate, and the contribution of $t_{m,h_o,h,\sigma}(m')$ to $e(m)$ would become marginal. The division of the integrand by $h^2(m_1 + m_2)$ normalizes e with respect to the height and the size of the template. We normalize with h^2 because the shape of the template is an average of all possible shapes, and thus does not need to match a peak, even in the absence of noise. This deviation grows linearly with the height, and we reduce its effect by the normalization. The task is now to find conditions for the error function e that characterizes the peaks. The first condition is $\partial e/\partial m = 0$, *i.e.*, e has a local minimum with respect to m. Usually, several local minima m_i of Eq. (2) are found in a neighborhood (Fig. 3), and we accept m_p

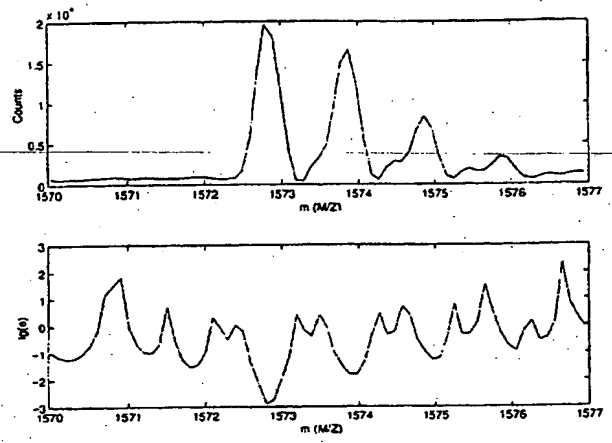

Figure 3. Error function (2) in the vicinity of a peak at 1572.835 Da. The lowest minimum corresponds to the monoisotopic peak. Note the logarithmic scale of e.

as a possible peak if $e(m_p) \leq e(m_i)$, $\forall m_i \in [m_p-(m_1 + m_2)/2, m_p + (m_1 + m_2)/2]$.

From these we select as real peaks those which satisfy the following conditions:

$$\left.\begin{array}{l} e(m_p) < e_{max} \\ h(m_p) > h_{min} n \end{array}\right\} \quad (3)$$

where e_{max} and h_{min} are thresholds and n is the estimation of the noise around the peak. Since the quality and height of the peaks vary, perfect values for e_{max} and h_{min} that are able to distinguish all 'true' peaks from noise do not exist. If the values are restrictive, *i.e.*, if e_{max} is small and h_{min} is large, no 'false' peaks are detected, but we also lose some 'true' ones. Conversely, by increasing e_{max} and decreasing h_{min}, more and more 'false' peaks appear (Fig. 6).

The values of e_{max} and h_{min} are linked to the values of the parameters one has to choose for the database search. For example, if the search values are restrictive, *i.e.*, the mass tolerance is low and the minimal number of peptide masses that must match is high, the values of e_{max} and h_{min} must be less restrictive in order for all 'true' peaks to be taken into consideration. The 'false' peaks do not change the result if they are not too abundant, because the probability that several of them match the same protein, thus giving rise to a high score for this irrelevant protein, is low. On the other hand, if the database search parameters are less restrictive, e_{max} and h_{min} may not allow many 'false' peaks, because the probability of a false match is now higher and the result may change qualitatively.

2.2.3 Implementation

The first step is to remove the trend because the height of the peaks is only defined relative to this trend, and because the numerical accuracy is improved. Ripley [23] describes several methods. The crucial point is the robustness of a method, *i.e.*, the trend need not follow localized deviations, like peaks. Here we choose a simple approach to attain a robust fit. The spectrum is split into several small windows s_i (width about 40 Da), and s_i^{med} and s_i^{low} are calculated in each window, where s_i^{med} and s_i^{low} are defined as follows: 50% of the values in s_i are larger than s_i^{med} and 50% are lower, while 95% of the values in s_i are larger than s_i^{low} and 5% are lower. Then we define the noise as $n_i = 2\,(s_i^{med}-s_i^{low})$. Finally, s_i^{med} and n_i are interpolated using cubic splines [24] to obtain the continuous trend $s^{med}(m)$ and noise $n(m)$ (Fig. 1).

If m and σ are given, calculating h_o and h is straightforward, thus minimizing Eq. (2). Hence we have to seek the minima e_i only in m and σ. Since this algorithm is also used for high throughput processes such as the molecular scanner [25], execution time is crucial. A direct evaluation of Eq. (2) was too slow, and it is therefore necessary to perform a fast first search to find starting points for a more extended search. This first search is done by fixing σ = 0.2 (this value neither produces too many minima, nor does it blur e too much; see Section 2.2.1) and evaluating Eq. (2) for masses where the signal exceeds the noise. Because the template t varies slowly with the mass, it does not have to be evaluated for each mass. We then calculate the minima of Eq. (2) and use the resulting masses as starting points for a more precise fit, where both m and σ vary. It is possible that two peptides have similar masses, so that their peaks overlap. In this case, the method described above may fail to detect both peaks. To solve this problem, all detected peaks are subtracted from the spectrum, and the algorithm is applied a second time.

2.3 Calibration

The TOF measured by a MALDI-TOF mass spectrometer can be affected with a significant error. After converting the TOF into the peptide mass [6], this can yield an error of up to 1 Da. However, most of that error can be corrected afterwards by a linear transformation:

$$m_{calib} = am + b \quad (4)$$

The coefficients a and b can be determined in three different ways. (i) They are defined externally. (ii) They are calculated using internal standards, *i.e.*, peptides with known masses that appear in the spectrum. This method works

well if internal standards are present and if they are detectable in the spectrum. Then it allows reducing the error to values smaller than 0.05 Da. (iii) They are calculated with a maximum likelihood method. This method is based on the fact that the mass distribution of peptides is not at all uniform. First, the distribution peaks at certain masses separated by 1 Da, and second, it drops with higher masses (Fig. 4) (for a detailed discussion see [22]). Let $P(m)\,\Delta m$ be the probability of finding a mass in $[m, m + \Delta m]$. For a set of peaks with masses m_i, a and b are chosen to maximize the total probability $\Sigma_i P(am_i + b)\,\Delta m$. This method is independent of internal standards, but it only works for initial errors that are smaller than 0.5 Da, making the error less than 0.2 Da in most cases.

2.4 Identification by peptide mass fingerprinting

2.4.1 Problems

Identification by peptide mass fingerprinting uses a set of experimental peptide masses obtained from the mass spectrum after peak detection, as well as information about the species, the isoelectric point or the molecular mass of the searched protein. These experimental masses are compared to a database of peptide masses, *i.e.*, a database of *in silico* digested proteins. The identification algorithm searches for the protein with the best match between its theoretical peptide masses and the experimental masses. Other attributes of the searched protein are also taken into consideration and matched to their corresponding values in the database. This method involves various problems that influence the quality of the identification.

First of all, we need to know which database to use, and how it has to be parsed. There are two approaches when using a database: either we search a database of protein sequences which is parsed linearly, each sequence being virtually digested to progressively determine the peptide masses, or we build an index of all possible peptide masses (sorted in ascending order) by an off-line digestion of a protein sequence database. In both cases we consider possible modifications and missed cleavages. The first method has the advantage of using less disk space, because everything is calculated on-line, and thus does not need to be stored. It also easily allows (by changing digestion rules) considering different enzymes for the digestion. Nevertheless, it could require a longer parsing time due to the fact that digestion operations, and especially all the combinations of modifications that could occur on peptides, have to be computed on-line. The second method has the advantage of retaining all possible peptide masses and therefore avoiding the combinatorial treatment of modifications during the search. Its drawbacks are the considerable additional space needed to store the index, as well as the time necessary to update it.

A second problem arises from the large number of parameters that influence the identification process. Indeed, as we have already seen, modifications and missed cleavages can occur and modify a protein's peptide masses. If we allow for each theoretical peptide to carry zero, one or several modifications and for the enzyme to miss 0, 1 or 2 cleavage sites, this strongly increases the number of

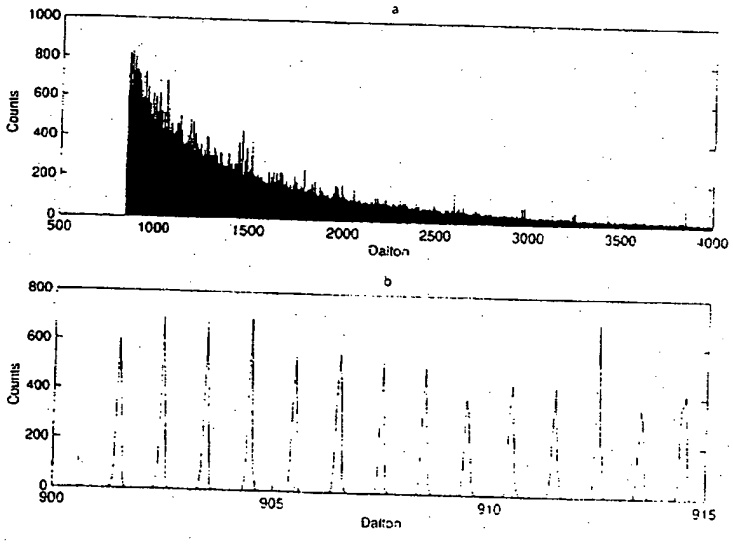

Figure 4. (a) Distribution of theoretical peptide masses obtained by a virtual digestion of all proteins in SWISS-PROT with trypsin (without missed cleavages or modifications). Note how the distribution drops for higher masses. (b) Detail of (a). The distribution is peaked at given masses.

theoretical peptide masses the program must consider. Also, identification algorithms use various thresholds that can appreciably modify the search results. Examples are the molecular weight range, the p*I* range, the minimum number of matches, the allowed difference between experimental and theoretical masses (mass tolerance), *etc.* For these reasons, the user should usually have an *a priori* idea of the experimental context, because an optimal choice of the parameter values will facilitate the interpretation of the results.

The third problem results from the way the resulting proteins are ranked by the identification algorithm. Depending on the parameters that have been selected for the search, the number of proteins in the database that match the experimental data can be very large. The program must therefore associate a score with each candidate protein, and thus allow the confidence in its match to be quantified.

2.4.2 Parameters

In order to choose an efficient method to handle the above-mentioned problems, the use of the parameters has been formalized. This has implications on the choice of the database structure and on the calculation of the score associated with each candidate protein. Parameters can be characterized in two ways. The first possibility concerns their effects on the quality and efficiency of the identification. The second possibility is linked to the level in the identification process at which a parameter intervenes.

When considering the first possibility one considers the fact that the parameters have two opposite effects during the search: an "exploratory" effect and a "filtering" effect. The exploratory effect allows an increase in the size of the search space, that is, an increase in the number of candidate proteins. Indeed, the first difficulty of the identification is to be sure to include the correct protein in the list of candidate proteins. Therefore the tolerance in the set of considered proteins and masses must be high enough to find the right protein. Parameters that are involved in this class of effects are: the type and number of modifications applied to proteins in the database, the maximum number of missed cleavages, the maximum distance between experimental and theoretical masses, the minimum number of matched peptides necessary for a protein to be selected, and the number of peaks returned by the peak detection program.

The second difficulty in the identification is to minimize the number of candidate proteins, in order to avoid losing important to efficiently filter the results and eliminate the least likely proteins from the list of candidates. Parameters with such a filtering effect include the species (to reduce the number of proteins to be considered), the molecular mass and the isoelectric point (to eliminate proteins whose values are too far from the experimental ones). Moreover, some of the parameters mentioned above for their exploratory effects, like the maximum distance of masses, the minimum number of matched peptides or the number of detected peaks, also have some filtering effects – depending on their thresholds.

The main difficulty consists in finding a compromise between these two aspects. On one hand, one wants to be sure to consider enough candidate proteins, therefore the exploratory effect has to be increased. On the other hand, one seeks to clearly identify the right protein and therefore has to filter the results. Depending on their exploratory or filtering nature, parameters may have a notable effect on the processing time needed for the identification. The more exploratory effects are used, the longer the search time will be. The sooner the filtering effects that are applied, the shorter it will be. The quality and efficiency of the identification will thus be highly dependent on the choices of the parameter values.

The second method of characterizing parameters is based on the levels at which parameters participate in the identification process. In the case of two-dimensional electrophoresis (2-DE), three levels can be considered. The first one, the "mass level", corresponds to the choice of mass used to match a protein. At this level, we want to characterize the degree of match between a mass found in the spectrum and the mass of a peptide of the search protein. The next level, the "protein level", consists of the identification of a protein at a given position in the 2-DE gel. Information from the mass level is coupled with information about the whole protein, in order to determine the best candidate protein. Finally, at the "contextual level", information about the two-dimensional environment (context) of the selected proteins from level 2 are taken into account to refine the identification at each position in the gel.

At the mass level, the first goal is to determine the quality of a peak, that is, to determine when a peak may be considered to be a "true" peak. For that purpose, parameters such as peak intensity, peak width or the peak's fit with a theoretical isotopic profile (see Section 2.2) can be used. A level of confidence is also defined for the match of an experimental mass with a theoretical mass in the protein database. This is achieved with the help of parameters such as the number and type of modifications, the num-

hydrophobicity value (GRAVY) [26] of the corresponding peptide. The latter estimates the probability of finding the peptide in the mass spectrum (the hydrophobicity value is important for the ability of a peptide to fly in the mass spectrometer).

At the second level, the protein level, we search in the set of candidate proteins for the protein showing the best correspondence with all information available from the gel and the spectrum. Values obtained at the mass level, as well as parameters describing the whole protein, can be used. Such parameters are the molecular mass and isoelectric point, but also the percentage of the protein sequence that is covered by peptides identified at level 1, or the standard deviation of the distance between theoretical and experimental peptide masses.

The contextual level allows an adjustment of the identifications obtained from the previous levels by taking the environment into account. For each position in the 2-DE gel where identification is attempted, the points in the neighborhood are considered. The distribution of the masses used for this identification, the distribution of the identified proteins, as well as of the parameters used in the previous steps are considered [25]. This method validates or invalidates certain parameters, thus altering the results of the previous levels. In this way, one can imagine an iterative method that gradually refines the identification by successive application of the three levels.

2.4.3 The algorithm

2.4.3.1 Initial choices

As we have seen, the choice of parameters as well as the point in time where they are used is decisive for the efficiency of the search. When choosing parameters the compromise between the sensitivity and selectivity of the search has to be considered. Moreover, the calculation of an identification score has to take into account the nature of parameters and the level at which they intervene. A preliminary study showed the importance of these parameters (see Section 3.2.1). We therefore developed a new identification tool based on the role and the relative importance of the various parameters, in order to determine a score allowing the best possible discrimination between the searched protein and the other candidate proteins. The algorithm limits the parameters with exploratory effects, while preserving enough sensitivity to be able to find most of the proteins. In that way, by strongly limiting the number of possible combinations arising from modification and missed cleavage parameters, one can obtain a fast and highly discriminant search algorithm, which does not produce too many candidates. The speed of the algorithm is also essential when it comes to automation of the process for large scale identification projects. Only one missed cleavage is allowed, as well as the following modifications: cysteine carboxymethylation, acrylamide adducts to cysteines and oxidized methionines. For these modifications, we permit only 0, 1 or all corresponding amino acids to be modified, in order to avoid too many combinations. Thus, the database can be parsed linearly and digested on-line, which avoids the use of a voluminous mass index. To improve efficiency, the database (SWISS-PROT and TrEMBL in FASTA format) have been split up into about 40 different sections, each of which contains the sequences of specific species or taxonomic category. A species tree was built that allows parsing only the part of the database corresponding to the user-specified organism or range of organisms. Finally, we consider the whole set of parameters with filtering effects, in the hope to modulate their usage and thus avoiding the effects of fixed thresholds which too radically eliminate interesting candidate proteins.

2.4.3.2 Definition of the score

The main difficulties in the definition of a score calculation are to determine the most important parameters, their relative weights and how to integrate the whole set of parameters into the score calculation. For this reason, we use the parameter levels defined above to determine a score calculation using their respective properties. Parameters of level 1, the mass level, serve to calculate a score of level 1, associated with each matching peptide. For a given protein, the contribution of the parameters of level 1 is the sum of the level 1 scores of its peptides. It can be seen as an extension of the notion of number of matches used by most of the existing identification tools that count the number of experimental masses matching theoretical peptide masses of the candidate proteins. The more identified masses a protein has in the mass spectrum, the higher is the confidence in its identification. While tools such as PepIdent and PeptideSearch assign a weight of either 0 or 1 to each peptide mass, depending on whether or not it is a match, our idea is that the weight associated with a peptide mass can be modified according to parameters of level 1. This gives an indication of the importance of a mass in the score calculation. We use four parameters at this level: the number of chemical modifications, the number of missed cleavages, the intensity of the corresponding peak in the mass spectrum, and the hydrophobicity coefficient. Then we calculate the first part of our score (S_1) by:

$$S_1 = \sum_{i=0}^{N} score_1(i)$$

where

$$score_1(a) = (coef_m)^{n_{mod}}(coef_c)^{n_{cleav}} coef_i(a) coef_h(a)$$

where N is the number of matched peptides, $score_1(a)$ the score of level 1 associated to peptide a, $coef_m$ the modification coefficient, $n_m(a)$ the number of modifications in peptide a, $coef_c$ the missed cleavage coefficient, $n_c(a)$ the number of missed cleavages in peptide a, $coef_i(a)$ the peak intensity coefficient of peptide a, and $coef_h(a)$ the hydrophobicity coefficient of peptide a. In this expression, the modification and missed cleavage coefficients are fixed for all peptides and all proteins. However, their importance is increased with the power of the number of modifications and missed cleavages that are present in the peptide, $coef_h(a)$ is proportional to the hydrophobicity of peptide a (the weaker the hydrophobicity, the higher $coef_h(a)$), while $coef_i(a)$ is proportional to the peak intensity of peptide a (the higher the peak intensity, the higher $coef_i(a)$).

Parameters of level 2 are used to compute coefficients that are then applied to the previously defined score. Indeed, at level 2, the parameters concern the whole protein, so they have to directly modify the value of the score associated to the protein. Four parameters are used at level 2: the molecular weight of the protein (M_r), its isoelectric point (pI), a coverage coefficient (the percentage of the protein sequence covered by the matched peptides) and a standard deviation of the distances between experimental and theoretical masses. The score of level 2 (S_2) is calculated as:

$$S_2 = \frac{1}{coe\ f_e}\ coe\ f_w\ coe\ f_p\ coe\ f_r \qquad (5)$$

where $coef_e$ is the standard deviation coefficient, $coef_w$ the molecular weight coefficient, $coef_p$ the isoelectric point coefficient and $coef_r$ the coverage coefficient. The criterion for considering the mass distance between experimental and theoretical masses for all matched masses is based on the fact that the more constant this distance is for all matched masses, the lesser is the likelihood that the matches happened randomly. This notion was refined to take into account the calibration error of the measuring device. Thus, we make a robust and iterative linear regression [27] upon all matched masses, and eliminate the masses that are too far from the regression line (which are more likely to be false matches). We then calculate the standard deviation of matched masses around this line. This regression is iterative as it is performed in several steps, each step eliminating the masses farthest from the regression straight line, the line then being recalculated based on the new set of masses. The iteration is stopped when no mass has been eliminated in the previous step, or when a given minimum number of masses is reached. The standard deviation calculated at this last step gives a hint of the correspondence between the mass alignment and the supposed spectrometer error. Moreover, one can expect that the linear regression compensates for some calibration errors occurring during the peak detection, thus stabilizing the overall algorithm.

M_r and pI coefficients are nonlinear. We define several thresholds for the distance between experimental and theoretical values of M_r and pI, and then associate a coefficient to each of these thresholds. The more the theoretical values move away from the experimental values, the weaker the coefficient is. The coverage coefficient is proportional to the percentage of the sequence that is covered, therefore the higher the percentage, the higher the coefficient.

Finally, the total score associated to a protein is given by the expression:

$$score = (S_1)^{\alpha} S_2 \qquad (6)$$

where α is a weight showing the importance of parameters of level 1 against those of level 2. Parameters of level 3 have not yet been taken into account, but they will be used within the scope of the "molecular scanner" [25].

2.4.3.3 The algorithm

The algorithm (1) used for the identification can be summarized as follows in a pseudoprogramming language.

2.5 The learning

The score calculation and the peak detection that we use involve many coefficients (some also requiring several thresholds) that are associated with the various parameters. These coefficients determine the relative importance of each parameter in the score calculation, in order to be able to best discriminate the right protein from the other candidate proteins. We use a learning algorithm to determine the coefficients and threshold values that allow the best discrimination. For this reason, the peak detection and identification parts of the algorithm have been unified to adjust all the parameters involved in the whole process, from spectrum analysis to identification. A genetic algorithm [28] has been applied to a training set of already identified proteins. This algorithm searches for the best coefficient values that allow the identification algorithm to identify the right protein, with its score being as distinct as possible from the scores of the following proteins in the ranked list of candidate proteins.

```
algorithm 1 identification
  enter_user_data()
  result_protein_list = ∅
  for all species ∈ species_list do
    for all databases ∈ databases_list do
      for all protein ∈ protein_list do
        (digestion of the protein and creation of the list of matched peptides)
        built_matched_peptide_list(peptide1_list)
        if number_match > match_threshold then
          (we take out from the peptide list those with a mass too far from the
          standard deviation)
          regression(peptide1_list, peptide2_list)
          score_calculation(protein, peptide2_list, score)
          add_result_protein_list(protein, score)
        end if
      end for
    end for
  end for
end for
sort_result_protein_list(result_protein_list)
```

2.5.1 The genetic algorithm

For the learning phase, 36 variables have been defined, representing all the parameters and thresholds needed to calculate the score. Among these 36 parameters, 33 have real values and only 3 have integer values. Therefore the parameters are coded as a vector (chromosome) or 36 real values (genes). We use a nonlinear mutation operator [29] for genetic algorithms working with real values. This operator decreases the mutation effect during generations, favoring the convergence of values associated with genes. We use a classical crossover operator and a readjusting fitness function operator [28], thus avoiding an all too rapid convergence of the algorithm. We have developed an extension to the classical genetic algorithm that uses two populations with different convergence levels, which optimizes the quality of the results. The population with a high convergence level contains 26 chromosomes and the one with a weak convergence level contains 44 chromosomes, each of them representing a set of particular parameters. We define a fitness function whose value characterizes how well our scoring function can discriminate the right protein. For the parameter values of each chromosome, we apply peak detection and identification algorithms to a subset of spectra from the training set. The results of these algorithms are then used to calculate the fitness value associated with each chromosome as follows:

$$value = \begin{cases} \frac{score_i}{score_1 - score_2} & \text{if } Rprot = prot_i \\ 0.5 - (position(Rprot) \cdot 0.05) & \text{else} \end{cases} \quad (7)$$

where $score_i$ is the score of the i^{th} protein from the list of results, *Rprot* the name of the searched protein, $prot_i$ the name of the i^{th} protein from the list of results, and *position*(*Rprot*) the position of the right protein in the list. The total fitness of the chromosome is the average of these values for the subset of spectra.

3 Results and discussion

3.1 Peak detection

Figure 5 shows a region of the SYP_ECOLI spectrum with peaks and their fits. Only a small fraction of the peaks could be interpreted as peptides of SYP_ECOLI; the other peaks may be due to impurities, protein fragments, or modifications. It reveals that small peaks may be important for identifying a protein. It is not clear *a priori* whether the simple classifier given by Eq. (3) is sufficient to separate 'false' peaks from 'true' ones. Therefore we performed a peak detection for the ten spectra used for testing the identification algorithm (Section 3.3) with threshold values of $e_{max} = 1$ and $h_{min} = 1$, which were not too restrictive. We plotted lg(*e*) *versus* $\lg(\frac{h}{n})$ (Fig. 6). This shows that there is a strong overlap between the 'true' and the 'false' peaks and the classifier (3) is not able to separate all of them. But the values of e_{max} and h_{min} given by the learning algorithm (Section 3.2) indicate that it is more important for the identification to consider all 'true' peaks, even if some 'false' ones mix in. Another result is the strong correlation between *e* and $\frac{h}{n}$, *i.e.*, the higher the peaks the better the fit. This is mainly due to the fact that we normalized the error function (2) with respect to the height.

3.2 The learning

3.2.1 Preliminary study

The influence of the effect of the main parameters upon the quality and speed of identification was studied. Studies have been performed by others, but without the use of experimentation [30]. For our study, the PepIdent tool was employed to identify a set of 20 known proteins, each time varying the values of the available parameters. A first result showed the dominant importance of the filtering parameters, especially the choice of a specific species and to a lesser degree the information about molecular weight and isoelectric point. Without these parameters, the correct protein was often lost among a very large set of candidate proteins. The analysis also highlighted the strong effect of modifications and missed cleavage parameters upon the number of generated candidates. Indeed, PepIdent takes into account annotations from SWISS-PROT entries and the chemical modifications that represent a huge combination of the number of different masses that are possible for a single peptide mass. Therefore, the quality of the results often deteriorates when one allows the whole set of modifications or, even worse, if one allows one or two missed cleavages. The analysis showed that certain proteins (when very few peptides from the protein are found in the spectrum and when

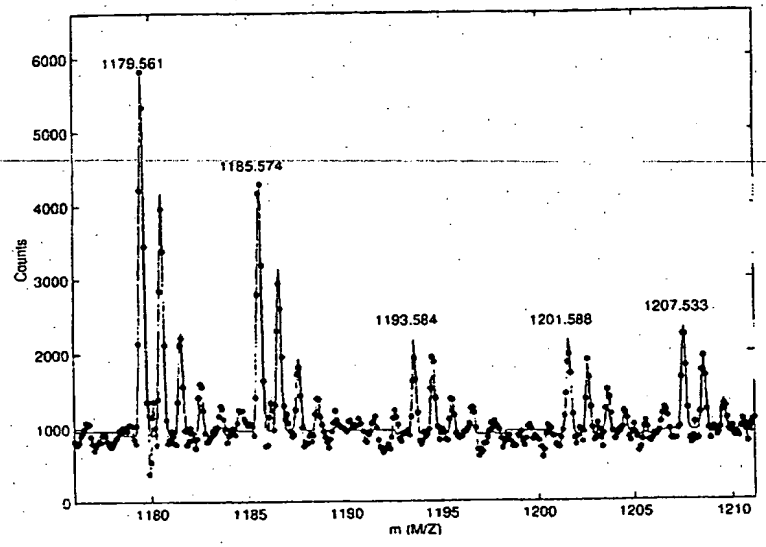

Figure 5. Detected peaks in the SYP_ECOLI spectrum (dots) and their fits (solid line). The values of the error function $e(m)$ were: $e(1179.561) = 0.37 \times 10^{-3}$, $e(1185.574) = 0.45 \times 10^{-3}$, $e(1193.584) = 2.63 \times 10^{-3}$, $e(1201.588) = 1.68 \times 10^{-3}$ and $e(1207.533) = 2.53 \times 10^{-3}$. Only the peaks at $m = 1185.574$ and $m = 1207.533$ match a peptide of SYP_ECOLI considering only chemical modifications in PepIdent and one missed cleavage.

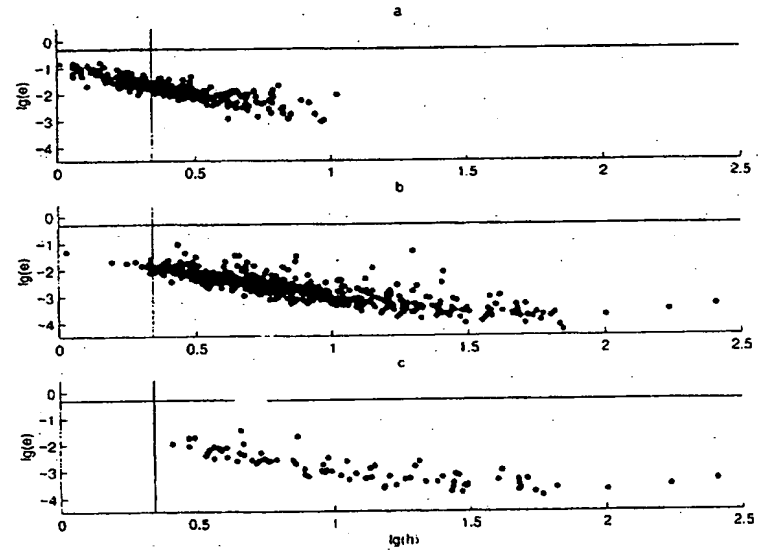

Figure 6. lg(e) *versus* lg(h/n) for the peaks of the ten test spectra with $e_{max} = 1$ and $h_{min} = 1$. (a) lg (e) *versus* lg(h/n) for the peaks that were considered as 'false' or at least uncertain after a visual examination of the spectra. (b) lg(e) *versus* lg(h/n) for the peaks that were considered as 'true'. (c) lg(e) *versus* lg(h/n) for the peaks that matched a peptide mass of the corresponding protein. The solid lines indicate the values of $e_{max} = 0.5$ and $h_{min} = 2.2$ obtained by the learning algorithm.

they are modified or incompletely digested) cannot be found without the use of at least one of these parameters.

3.2.2 Genetic algorithm

We selected a set of 91 proteins with known identification (identified with at least two methods, including peptide mass fingerprinting, microsequencing, gel matching and amino acid composition analysis) as a training set. We carried out several learning phases, gradually increasing the number of parameters, the number of proteins in the training set, and varying certein parameters of the genetic algorithm that influence its convergence level. Each application of the peak detection and identification algorithm takes about 1 min (on an Ultra Sparc Station 5, Sun Microsystems Inc.), so it is not possible to test the whole set of 91 spectra for each chromosome. Instead, we randomly chose 20 spectra for each chromosome and defined the fitness to be the average of their score. We can estimate the execution time of our learning algorithm for 100 generations: $100 \times 70 \times 20 = 140\,000$ min (about

100 days). Due to the time needed for a complete execution of the algorithm, we currently only have partial results. The version of our learning algorithm that uses the whole set of 36 parameters presented in this article and the 91 spectra of the data set was still running when the article was submitted. However, we present here the results obtained by the previous execution, using 29 parameters and 58 spectra during 24 generations.

The parameters of this execution are: the two parameters of peak detection e_{max} and h_{min}, the parameter *calibFact* which influences the calibration weight in the peak detection, *minMatch* which gives the minimum number of matches necessary to consider a protein as a candidate, *deltaMass*, the maximum allowed tolerance of masses, the two coefficients *coefReg*1 and *coefReg2* which are the thresholds for eliminating masses with linear regression, the coefficient *coefMiss* applied to a missed cleavage, the coefficient *coefModif* applied to a modification, the four coefficients *MWcoef*1 to *MWcoef*4 applied to the deviation of molecular masses and associated to the four threshold parameters *MWthres*1 to *MWthres*4 for the deviation of molecular masses, the four coefficients *PIcoef*1 to *PIcoef*4 applied to the deviation of isoelectric points and associated to the five threshold parameters *PIthres*1 to *PIthres*5 for the deviation of isoelectric points, the two parameters *nbMatchThres*1 and *nbMatchThres2* which determine the threshold at which the iteration on the linear regression is stopped, and *power* (α), the weight applied to the parameters of level 1 against those of level 2.

Figure 7 shows the results obtained with these parameters. The first one, *fitness*, shows the algorithm convergence with three curves. The lowest one corresponds to the average fitness of the population with a weak convergence level, and the following one to the one with a high convergence level. We can see very good convergence of the population with a high convergence level, with a maximum average fitness of 0.8123 for 26 chromosomes, corresponding to 520 identifications. The uppermost curve gives the value of the best chromosome for each generation, with a maximum at 0.9315 (20 identifications). These chromosomes are used to determine the best parameters for the identification algorithm at the end of the learning step. Therefore we present the evolution of the parameters of the best chromosomes in the following graphs. Note, however, that these results depend on the data used for the learning. The lack of variety in the data for a given parameter can cause a bias in the obtained results. In the future it will be important to repeat the learning with a larger set of data and as much diversity as possible.

The parameters that present the clearest results are h_{min}, *minMatch* and *coefMiss*. The high convergence of h_{min} shows that this parameter is important for the identification. Its value implies the importance of considering all 'true' peaks even if some fictitious ones mix in. Because of the strong correlation between h_{min} and e_{max}, peaks are primarily selected by h_{min}, and e_{max} does not intervene. The results of *minMatch* clearly show that the exploratory effects of this parameter are more important than its filtering effects, thus avoiding the loss of small proteins among the candidate proteins. The very weak values of the coefficient associated with the missed cleavage prove that the combination due to the use of missed cleavages implies such a huge increase of false matches, that the weight associated to peptides with missed cleavages must be drastically reduced. This corresponds to cases with very good digestion. We can also deduce that for an algorithm that does not incorporate a penalizing factor for missed cleavage, it is preferable in case of good digestion not to use the possibility to allow for missed cleavages at all.

Some other results are also rather clear, such as the high value of *deltaMass* that gives an important exploratory effect of level 1 in the matching of peptides that can be compensated for by the filtering effect of the linear regression used only at level 2. The high values of parameter *coefReg*1 show that it is preferable to eliminate masses only when they are far enough from the line of the linear regression (level 1). In any case their weight is lowered by the value of the standard deviation (level 2). The values of parameter *power* can imply that the weight to give to the parameters of level 1, compared to those of level 2, must be higher than what was allowed in this experience. A larger variation interval has therefore been permitted for this parameter in the new experimentation currently under way. The high variation of values of the *calibFact* parameter confirms the limited role of the calibration, due to the use of linear regression.

Due to the small number of generations calculated, the variability of the other parameters cannot be clearly explained yet. They may not have converged at the time of writing, but one can probably say that they are less discriminant than those presented before. One more global conclusion is that the division of the score calculation into levels that allow considering parameters at various steps of the search process is very important to resolve conflicts between exploratory and filtering effects. The exploratory effects can be efficiently used (if they are not too costly, as is the case for missed cleavages), if later strong enough filtering effects are present to compensate for their effects. Thus, we obtain a search algorithm that studies a maximum of candidate proteins, while preserv-

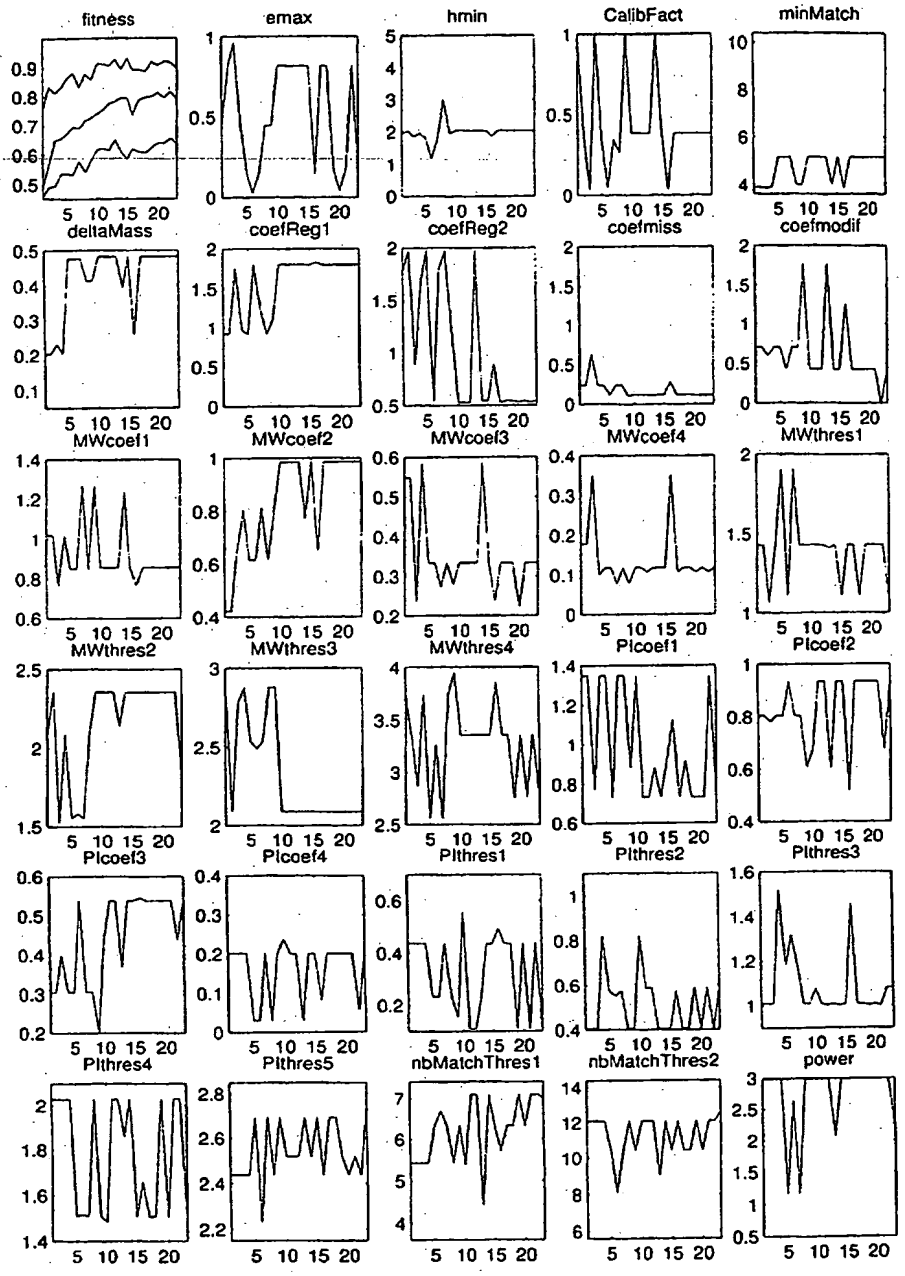

Figure 7. Learning of parameters. *X* axes correspond to the number of generations and *y* axes correspond to the values of parameters.

ing sufficient discrimination power to bring out the right protein.

3.3 Comparison

We have developed an identification tool, PeptIdent2, based on the method presented in this paper. In order to validate the method, we have undertaken a comparative study of the quality of protein identification obtained by several identification tools. We compare the results of PeptIdent2 with those of PeptIdent, Mowse, ProFound, PeptideSearch and MS-Fit (see Section 1.2). For this, we took a set of ten mass spectra of proteins whose identifications have been confirmed by microsequencing. For each protein, we show in Table 1 and 2 the identification result for each of the algorithms. For each identification

Table 1. Raw comparison of identification tools

	PeptIdent2	PeptIdent	Mowse	ProFound	PeptideSearch	MS-Fit
P56480	1(335.5)I2(9.38)	1(15)I2(7)	1(3.59^{-10})I8(1.42^{-7})	1(2.9^{-1})I5(8.1^{-4})	1(13)I3(12)	1(2.39^{+5})I2(126)
P02088	1(205.2)I2(23.9)	1(22)I9(9_{4ex})	1(1.5^{-6})I2(3.99^{+5})	1(5.0^{-1}$_{2ex}$)I3(8.1^{-4})	1(7_{2ex})I3(6)	1(4.83^{+6})I2(1.62^{-3})
P01942	1(9.85)I2(2.76)	–	–	1(3.4^{-1})I3(1.7^{-1})	–	1(65.3)I3(49.8)
P17742	1(9.81)I2(0.29)	1(5)I5(4_{3ex})	–	–	1(4_{4ex})I5(3)	1(189)I2(19.9)
P43024	1(16.21)I2(12.72)	1(18)I18(6_{3ex})	1(3.67^{+5})I8	1(3.4^{-1})I14(3.3^{-4})	1(5_{8ex})I9(4)	–
P10639	1(2.31)I2(1.72)	–	–	–	–	–
P12787	1(41.54)I2(10.61)	–	– (2.88^{-4})	1(4.9^{-1}$_{2ex}$)I3(4.3^{-3})	1(8)I3(7_{2ex})	–
P27773	1(209.69)I2(11.82)	1(12_{2ex})I2(11)	–	1(3.2^{-1})I4(1.6^{-1})	1(12)I2(11)	1(7.15^{+4})I2(394)
P27773	1(557.6)I2(121.2)	1(27)I17(15_{4ex})	–	1(1.0)I2(1.9^{-5})	–	1(5.81^{+4})I2(2.96^{-4})
P38647	1(737.0)I2(84.62)	1(19)I5(13_{2ex})	1(1.18^{-12})I4(1.58^{-11})	1(2.1^{-1}$_{2ex}$)I3(1.9^{-1})	–	1(1.99^{+6})I2(8.23^{-3})

Table 2. Comparison of identification tools after user analysis

	PeptIdent2	PeptIdent	Mowse	ProFound	PeptideSearch	MS-Fit
P56480	1(335.5)I2(9.38)	1(15)I2(6)	1(1.42^{-7})I2(9.2^{+4})	1(8.1^{-4})I2(1.1^{-21})	1(12)I2(5)	1(2.39^{+5})I2(126)
P02088	1(205.2)I2(23.9)	1(10)I2(5)	1(1.5^{-6})I2(1.81^{-5})	1(5.0^{-1}$_{2ex}$)I3(8.1^{-4})	1(7_{2ex})I3(5)	1(4.83^{-6})I2(1.62^{-3})
P01942	1(9.85)I2(1.11)	1(5)I2(4_{4ex})	–	1(3.4^{-1})I2(1.7^{-1})	–	1(65.3)I3(49.8)
P17742	1(9.81)I2(0.19)	1(4)I2(3)	–	–	1(4_{4ex})I5(3)	1(189)I2(19.9)
P43024	1(16.21)I2(12.72)	1(6)I2(4)	1(4.45^{+4})I5(2.88^{+4})	1(1.1^{-1})I12(3.3^{-4})	1(5_{6ex})I9(4)	–
P10639	1(2.31)I2(1.72)	1(3_{7ex})	–	–	–	–
P12787	1(41.54)I2(10.61)	1(8)I2(7)	–	1(4.9^{-1}$_{2ex}$)I3(4.3^{-3})	1(7)I2(5)	–
P27773	1(209.69)I2(11.0_{2ex})	1(12)I2(9_{2ex})	–	1(2.3^{-1})I2(1.6^{-1})	1(12)I2(11)	1(7.15^{-4})I2(394)
P27773	1(557.6)I2(121.2)	1(14_{ex})I2(13_{2ex})	–	1(1.0)I2(1.9^{-5})	–	1(5.81^{+4})I2(2.96^{-4})
P38647	1(737.0)I2(23.02)	1(15)I2(10_{5ex})	1(1.58^{-11})	1(2.1^{-1}$_{2ex}$)I3(9.9^{-2})	–	1(1.99^{+6})I2(8.23^{-3})

we also give the score value (between parentheses) of the first candidate protein followed by either the score value of the second candicate protein (if the first one is the right protein) or, otherwise, the rank and, in parentheses, the score value of the right protein. The right protein is always displayed in bold type. The notation *Xex* means that the score of the corresponding protein is equal to that of *X* other proteins, the algorithm not being able to give a clear discrimination. Finally, we note '– ' if the right protein was not found among the first twenty candidate proteins.

For these experiments, the parameters used in all identification programs were identical, if these parameters were available for each respective tool. The selected species was mouse, the allowed M_r variabiity was ± 50%, the allowed p*I* variability was ± 1, the minimum number of matched masses was 3, the maximal tolerance for masses was 0.3 Da, at most one missed cleavage was allowed and the modifications taken into account were cysteine carboxymethylation and oxidized methionines. Table 1 gives "raw" results, that is, without user interpretation. In this table the databases used were SWISS-PROT and TrEMBL for PeptIdent and PeptIdent2, SWISS-PROT for MS-Fit, OWL for Mowse, nrdb for PeptideSearch and NCBInr for ProFound. Table 2 gives the results after a first analysis by an expert user in our laboratory, in particular to remove the proteins with species that did not correspond to the search, as Mowse, ProFound and PeptideSearch do not narrow down the search based on species. TrEMBL database was also removed for PeptIdent and PeptIdent2 tools to have a better comparison with MS-Fit, which cannot use TrEMBL.

The first thing we notice is the good identification obtained by PeptIdent2 in both tables. In the second table, the right protein was identified in the first place in 9 out of 10 cases, and with a large score discrimination (at least five-fold) in 6 out of 10 cases. The only protein that was not correctly identified was P10639, which ranked second in the list of results, with a score quite close to the one of the first protein. No other identification program correctly identified this protein, except for PeptIdent which put it in first place with six other proteins of identical score. PeptIdent globally allowed good identification when the TrEMBL database was not used, but with a much less clear discrimination than PeptIdent2, and many proteins

were attributed identical scores. The other programs obtained variable results, the best being ProFound and MS-Fit. Note that programs that do not select a species (Mowse, ProFound and PeptideSearch) give result lists that are much larger, thus requiring a much larger manual analysis time to select the right protein in the list. Moreover, in this case, the risk is higher that the right protein does not appear at all on the list of results. One can also note that programs that use only the number of matched peptides (PepIdent and PeptideSearch) as their score have a much weaker discrimination power than the others and more often find proteins with identical scores, making the interpretation of the results by the user more difficult.

We are now following this comparative study with a second one on a larger set of proteins and with various species in order to obtain better validation of the comparison. To preserve a maximum reliability in the comparison results, we plan to use only experiments in which MS identification has been at least confirmed by microsequencing.

4 Concluding remarks

Protein identification and characterization is one of the most essential tasks performed in proteome research. The currently most widely used identification method compares the masses obtained from an MS spectrum of an enzymatically digested protein with the theoretical masses of proteins contained in an *in silico* digested protein sequence database. The precise determination of the peptide masses in the spectra, and a highly discriminating mass comparison algorithm are therefore the keys to the accurate identification of proteins. We have developed a new tool to identify proteins from their peptide mass fingerprints. It comprises a fast and precise peak detection algorithm, as well as a new mass comparison and identification program, which is based on an advanced scoring method, both procedures being validated by an automatic learning algorithm. The analysis of the thresholds associated with the peak detection has revealed that it is preferable to be little selective in the choice of peaks in the mass spectrum in order to avoid the loss of apparently fictitious peaks that might eventually appear to be useful, provided the identification algorithm is able to discriminate 'false' peaks from real ones. Our identification algorithm has proven to be robust enough in this respect. Also, the learning procedure has confirmed the advantage of a scoring scheme based on the balance between exploration and discrimination. This has led to a significant gain in the discrimination of the correct protein, in comparison to other identification algorithms.

This work is now being extended by the development of a new version of the learning algorithm that will be able to classify the proteins in the learning set simultaneously with the calculation of the parameter weights. This will determine several subsets of the parameter space, thus allowing an optimal discrimination of the scores. The goal is to determine several sets of parameters that will optimally discriminate the scores, no longer for all proteins, but rather for one subset of proteins that corresponds to a specific value of one of the experimental parameters (species, M_r, pI, etc.). Our score calculation will also be extended at the contextual level within the frame of the development of our molecular scanner. In addition, a new intermediate level, the "correlation level", will be introduced between the protein and the contextual level, which will consider information from several experiments carried out with different experimental conditions producing several fingerprints of the same sample. The correlation of these data will then validate the information obtained from the preceding levels. We thus expect to further improve the efficiency of our protein identification method.

This work was supported by the Swiss National Fund for Scientific Research (grant 31-52974.97) and the Helmut Horten Foundation. The authors would like to thank Dr. Eva Jung for useful discussions and Luisa Tonella, Gerald Rosselat, Salvo Paesano and Abderrahim Karmime for preparing the samples and testing the software.

Received July 30, 1999

5 References

[1] Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., Williams, K. L., *Biotechnol. Genet. Eng. Rev.* 1995, *13*, 19–50.
[2] Edman, P., Begg, G., *Eur. J. Biochem.* 1967, *1*, 80–91.
[3] Karas, M., Bahr, U., Giessmann, U., *Mass Spectrom. Rev.* 1991, *10*, 335–357.
[4] Muddiman, D. C., Gusev, A. I., Hercules, D. M., *Mass Spectrom. Rev.* 1995, *14*, 383–429.
[5] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., *Mass Spectrom. Rev.* 1990, *9*, 37–70.
[6] Cotter, R., *Time-of-Flight Mass Spectrometry, ASC Professional Reference Books*, Washington, DC 1997.
[7] Koster, C., Kahr, M. S., Castero, J. A., Wilkins, C. L., *Mass Spectrom. Rev.* 1992, *11*, 495.
[8] Cooks, R. G., Hoke, S. H., Morand, K. L., Lammert, S. A., *Int. J. Mass Spectrom. Ion. Proc.* 1992, *118*, 1–36.
[9] Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F., *Proteome Research: New Frontiers in Functional*

[11] Bairoch, A., Apweiler, R., *Nucleic Acids Res.* 1999, *27*, 49–54.

[12] Pappin, D. J. C., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[13] Zhang, W., Chait, B. T., *The 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, 1995.

[14] Gonnet, G. H., *A Tutorial Introduction to Computational Biochemistry Using Darwin*, Technical Report, E. T. H. Zürich, Switzerland, November 1992.

[15] Laemmli, U. K., *Nature* 1970, *277*, 680–685.

[16] Hochstrasser, D. F., Harrington, M. G., Hochstrasser, A. C., Miller, M. J., *Anal. Biochem.* 1988, *173*, 424–435.

[17] Cotter, R., *Time-of-Flight Mass Spectrometry*, Chapter 10, ASC Professional Reference Books, Washington, DC 1997.

[18] Klimowski, R. J., Venkataraghavan, R., McLafferty, F. W., Delany, E. B., *Organ. Mass Spectrom.* 1970, *4*, 17–39.

[19] Sukharev, Y. N., Nekrasov, Y. S., *Organ. Mass Spectrom.* 1976, *11*, 1232–1238.

[20] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C*, Chapter 15, Cambridge University Press, Cambridge 1992.

[21] Canny, J., *IEEE Trans. Pattern Anal. Machine Intell. PAMI* 1986, *8*, 679–697.

[22] Gay, S., Binz, P.-A., Hochstrasser, D. F., Appel, R. D., *Electrophoresis* 1999, *20*, 3527–3534.

[23] Ripley, B. D., *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge 1996.

[24] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C*, Chapter 3, Cambridge University Press, Cambridge 1992.

[25] Binz, P. A., Müller, M., Walther, D., Bienvenut, W. V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M. R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmime, A., Bairoch, A., Sanchez, J.-C. Appel, R., Hochstrasser, D. F., *Anal. Chem.* 1999, in press.

[26] Kyte, J., Doolittle, R. F., *J. Mol. Biol.* 1982, *157*, 105–132.

[27] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C*, Cambridge University Press, Cambridge 1995.

[28] Goldberg, D. E., *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA 1989.

[29] Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, Berlin 1996.

[30] Mann, M., *Microcharacterization of Proteins*, Chapter VI.2, Wiley-VCH, Weinheim 1994.